An Attempt to Automate the Process of Source Evaluation

Sonal Aggarwal[1] and Herre Van Oostendorp[2]

[1] International Institute of Information Technology, Hyderabad, India
Email: sonalaggarwal1@gmail.com

[2] Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
Email: herre@cs.uu.nl

Abstract — **Credibility of a web-based document is an important concern, when a large number of documents is available on internet for a given subject. In this paper, various criteria that affect the credibility of a document are explored. An attempt is made to automate the process of assigning a credibility score to a web-based document. Presently the prototype of the tool developed is restricted to only four criteria – type of website, date of update, sentiment analysis and a pre-defined Google page rank. Also a separate module for checking "link integrity" of a website is developed. To obtain empirical validity of the tool, a pilot study is conducted which collects credibility scoring for a set of websites by human judges. The correlation between the scores given by human judges and the scores obtained by the tool developed is low. The possible reasons for the low correlation are firstly, the tool is restricted to only four criteria, and secondly, subjects themselves had no agreement. Apparently they judged the website on different criteria, and not weighted overall. Further enhancements to the work done in this paper can be of great use to a novice user, who wishes to search a reliable web-based document on any specific topic. This can be done by including all criteria (discussed in this paper) for calculating the credibility score of a website.**

Index Terms — **credibility score; source evaluation; automation; web-based multiple documents**

I. INTRODUCTION

A. Need for Source Evaluation of web-based Documents

The number of web-based documents available for a given subject is tremendous. So it becomes necessary to evaluate the sources of these documents in order to choose the most appropriate documents. Studies conducted by Metzger et al. [1] indicate that particularly college students rely heavily on the web for both general and academic information, but verify the information very rarely. Another study by Walraven et al. [2] also indicates that students do not frequently explicitly evaluate 'sources' and 'information' with respect to a web-based document. Empirical psychological research shows that adequately evaluating the credibility of sources is important: students who do this well or assess reliability and use source characteristics achieve better comprehension of the content [3]. Research by Amin et al. [4] demonstrates that providing a novice user with credibility scores of web-pages boosts the confidence level in the selection of information, though it does not make search more efficient. Case studies by [5] emphasize on critical analysis of internet and scholarly sources by undergraduates, as they are unable to discriminate between credible and non-credible sources.

In this paper we try to automatically assign credibility indices to web-based documents, based on different criteria and report back to the users.

B. Various criteria of source evaluation

The different criteria based on which we can evaluate a web-based source are as follows: *Type* of website. A website can be an educational website (.edu, .ac), government website (.gov), a commercial website (.com), organizational website (.org) or other website. Depending on the type, a website can be more or less reliable; *Date* of the web-site, when it was last updated, also affects the credibility of a web-based document; A web-document being a *primary or secondary source,* also decides the credibility of the document [6]; Availability of *contact information* (address and/or email id) of the owner of the web-document; Analyzing the *link integrity* of the website [7]. A website with balanced internal and external links is more credible. Also it should not have any broken link; Analyzing the header/ footer of the web-site for any *affiliation* (if available) [8]; *Completeness* and *accuracy* of the information; Author's *expertise* in the subject [9]; Author's *opinion* biased or un-biased. Analyzing the sentiment (positive, negative or neutral) of a website can determine author's opinion; Author's *connection* to the *source* of publication; *connection* to the intended audience [6]; Author's *point of view* is objective and impartial or not [6]; Author's *credentials* like institutional affiliation (where he or she works), educational background, past writings, or experience. [6]; *Purpose* of the web-page (somewhat reflected by the type of website); *Interactivity, Usability* of the web-site; *Structure* of the web-site in terms of graphics and text are appropriate or not [7]; *Quality* of information on the web-site: elementary, technical, or advanced; *"Tone"* of the webpage: ironic, humorous, exaggerated or overblown arguments [10]; Determining if advertising and informational content are being supplied by the *same person or organization*; If so, advertising likely to bias informational content [8]; Determining any *software requirements* that may limit access to web information [11]; Is the *web-site better* than the other. If so why; *Ranking of the web-source with Google.* Google has a patented Page-Ranking technology that can also be criterion of source evaluation [12]; *Domain Experts' view* on the credibility of any web-document. As according to Amin [13], experts develop different strategies while seeking information from internet. These strategies can be helpful to a novice user. An expert can assess the credibility of a web document on the basis of two dimensions of credibility

18

– trustworthiness (well-intentioned, unbiased) and expertise (knowledgeable, competent) [14]. C. Multi-Criteria Decision Analysis (MCDA) When a decision is affected by more than one criterion, a Multi-criteria decision analysis is required. One of the methods for doing this analysis is "**P**otentially **A**ll **P**airwise **R**an**K**ings of all possible **A**lternatives" (PAPRIKA), in which a pair-wise ranking of all possible alternatives is done, so as to identify all dominating and un-dominating pairs. The dominating pairs are given more priority [15]. In this paper we are using this method to define initial weights to available criteria for defining and computing credibility indices.

II. Hypothesis

It is assumed that we can automate the process of assigning credibility indices to web-based documents based on various criteria for source evaluation. Presently to compute the credibility index of a website we are considering only four criteria: type of website, date of update, sentiment analysis and Google page Rank. To examine whether we have achieved our goal we will compare the performance of the tool regarding the credibility index with the rating of human judges.

III. Method

A. Design

A Prototype of the tool is designed which takes Google Search results/Wikipedia external links for a given topic. And assigns a credibility score to each web-document based on different criteria. The weights (Table 1) are assigned using 1000 Minds [16], decision making software which implements PAPRIKA method.

TABLE I
WEIGHTS ASSIGNED TO EACH CATEGORY

Criterion	Category	%	Points
Type of Website	Gov	37.6%	100
	Edu, org	30.3%	80.58
	Info, net	5.6%	14.89
	Com	0.9%	2.39
	Others	0%	0
Date of Update	Less than 1 year	13.2%	100
	>1yr & <5yr	6.4%	48.48
	>5yr	0.4%	3.03
	Not available	0%	0
Sentiment	Neutral	21.4%	100
	Positive	7.7%	35.98
	Negative	0%	0
	Not available	0%	0
Google Rank	9-10	27.8%	100
	7-8	24.4%	87.7
	5-6	6.8%	24.46
	3-4	2.6%	9.35
	1-2	0%	0

Based on the above weights, a credibility score is assigned to any website. For example if an organizational website which is recently updated, having positive bias and rated as 7 by Google ranking, will be given the following score : (80.58+100+35.98+87.7)/4 = 76.065 The results (search results with credibility scores) are displayed in a tabular form (see Figure 1). User is allowed to edit any score or to download all the scores.   Along with this, a separate module for checking link integrity is developed, which checks the link integrity of a website address entered by the user.

A. Materials & Equipments

The software tools used in the development of the prototype source evaluation tool are: Python [17] and Py2exe [18]; Pattern [19]; AlchemyAPI [20]; .Net Framework for web interface; Microsoft Access.

B. Procedure

A Python script is developed, which gets search results from Google or external links of Wikipedia page for a given topic using Pattern, does sentiment analysis using AlchemyAPI, obtains Google Rank for which url from http://webinfodb.net/a/pr.php?url=<website_url>, obtains type of website by analyzing the url, calculates credibility score for each search URL based on four criteria and writes the result to a database file. The script uses Multi-threading concept so that it can make parallel calls to different URLs. And it is converted into an executable program using Py2exe. Another Python script is developed which checks the 'link Integrity' of a given URL It gives output as total number of 'internal links', 'external links' and 'broken links'. This script makes call to each link available on the given web-page. If the link is not opening – it is considered as broken link. While if the link is '#' it is a self-link. If a link to a web-page in the same domain exists then it is said to be an internal-link. And if it goes to another domain it is an external-link. A web-interface is developed which uses the above python scripts for back-end processing.

III RESULTS

A. Output Source Evaluation Tool

The output of the tool when given a search topic "Tourism in India" is shown in Figure 1. The user is also allowed to edit or download results as a database file.

B. Empirical Vaility

A small group of seven people was asked to give credibility scores (1-10) for nine websites. The pilot study was conducted by sending an email to the group. The participants were provided with the URLs of nine websites with a small introduction about the factors affecting the credibility. With a concern of not biasing the opinion of the group, the introduction part (in the form of question and answer) was made brief. The results were matched with the scores provided by the developed tool. The scores given by the participants and the tool are as shown in table 2. Based on the mean of scores obtained by the participants the correlation coefficient of 0.484, $p < .19$ is obtained.

TABLE II
SCORES GIVEN BY PARTICIPANTS AND THE TOOL

Website URL	Individual Rating of participants (out of 10)	Score by tool (out of 100)
http://tourism.gov.in/	5,5,5,7,7,2,10	65
http://india.gov.in/overseas/visit_india/medical_india.php	6,4,6,7,9,9,3	56
http://www.incredibleindia.org/	2,7,8,9,9,1,9	51
http://www.jaipur.org.uk/	7,4,8,6,8,10,6	35
http://www.travelagewest.com/travel/asia/Info/Medical-Tourism-to-India/	9,5,4,3,7,6,2	25
http://www.mumbai.org.uk/	1,4,7,6,6,8,5	32
http://www.indiaholiday.org/india-tourism.html	10,5,7,5,8,3,8	29
http://www.agritourism.in/	4,4,6,2,6,7,4	14
http://www.tourism-of-india.com/	3,6,7,4,8,4,7	16

V. CONCLUSIONS & DISCUSSION

The prototype of the tool developed uses only four criteria for assigning credibility score to the website. The pilot study conducted to obtain empirical validity of the tool, gives a correlation with a low value (.48). The possible causes of this low correlation value can be summarized as follows: The tool is restricted only to four criteria. If all criteria (discussed in section I (B)) are included to obtain the credibility score, the system may give better results. Only then a higher correlation with subjects' judgments will be found; The weights assigned by the subjects had no agreement in themselves. Apparently they judged the website on different criteria, and not weighted overall. The work done in this paper can be enhanced further, by including all the criteria of source evaluation into an automated system. Some of the criteria may not be possible to evaluate automatically. We may use a database having meta-data given by experts in that case. Development of this system will be helpful to a novice user, giving him/her a level of confidence on the reliability of any specific web-document. Also integration of this Source Evaluation tool to existing automatic multi-document summarizers will help to achieve a better summarization.
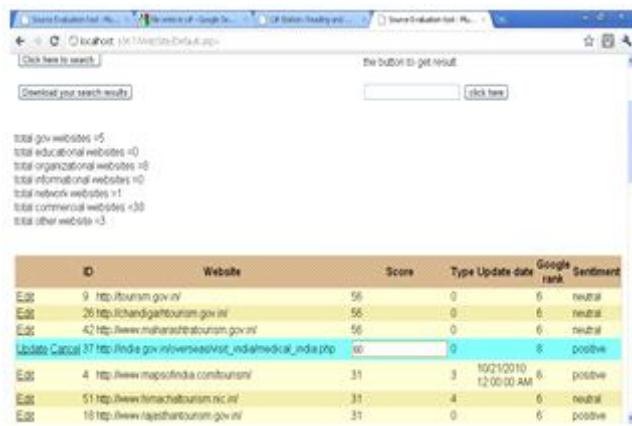

Figure1. Output of the tool for a searched topic

REFERENCES

[1] M.J. Metzger, A.J. Flanagin & L. Zwarun, 2003. "College student Web use, perceptions of information credibility, and verification behavior," *Computers & Education*, volume 41, pp. 271–290.

[2] Amber Walraven, Saskia Brand-Gruwel & Henny Boshuizen, "How students evaluate information and sources when searching the World Wide Web for information", *Computers & Education*, 2009, 52, 234-246.

[3] Braten, I., Strømso, HI, & Britt, MA (2009), "Trust matters: Examining the role of source evaluation in students' construction of meaning Within and across multiple text," *Reading Research Quarterly*, 44, 6-28.

[4] A. Amin, J. Zhang, H. Cramer, L. Hardman & V. Evers, "The effects of source credibility ratings in a cultural heritage information aggregator," *in Proceedings of the 3rd workshop on information credibility on the web*, New York, NY, USA: ACM, 2009, pp. 35–42.

[5] Calkins, S., & Kelley, M. R. (2007), "Evaluating Internet and scholarly sources across the disciplines: Two case studies," *College Teaching*, 55 (4), 151-156.

[6] Joan Ormondroyd , "*Critically Analyzing Information Sources*," September 2009, from: http://olinuris.library.cornell.edu/ref/research/skill26.htm

[7] Esther Grassian, the UCLA Library, "*Thinking Critically about World Wide Web Resources*," June 1995, from: http://www2.library.ucla.edu/libraries/college/11605_12337.cfm

[8] Beck, Susan, "The Good, The Bad & The Ugly: or, Why It's a Good Idea to Evaluate Web Sources*,*" 1997, from: *http://lib.nmsu.edu/instruction/evalcrit.html*

[9] Laura Cohen, Trudi E. Jacobson , "Evaluating Web Content*"*, March 2009, from: *http://library.albany.edu/usered/eval/evalweb/EvaluatingWebContent.pdf*

[10] UC Berkeley - Teaching Library Internet Workshops, "Finding Information on the Internet: A Tutorial*,*" Copyright 2010, from: *http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html*

[11] Jan Alexander & Marsha Tate, "The Web as a Research Tool: Evaluation Techniques," 1996-1998, from: *http://www.mediaawareness.ca/english/resources/educational/teaching_backgrounders/internet/web_as_research_tool.cfm*

[12] Google Technology, from: http://www.google.com/corporate/tech.html

[13] A. Amin, J. van Ossenbruggen, L. Hardman & A. van Nispen, "Understanding cultural heritage experts' information seeking needs*,*" In Proc. JCDL '08, pages 39{47, New York, NY, USA, 2008. ACM.

[14] B. J. Fogg and H. Tseng, "The elements of computer credibility," *In Proc. CHI '99*, pp. 80-87, ACM.

[15] An Article on Multi-criteria decision analysis, from: *http://en.wikipedia.org/wiki/Multi-criteria_decision_analysis*

[16] 1000Minds, advanced decision-support software, from: *http://www.1000minds.com*

[17] Python, from: *http://www.python.org*

[18] Py2exe, from: *http://www.py2exe.org*

[19] De Smedt, T. & Daelemans, W. (2011), Pattern version 1.0, retrieved January 2011, from: *http://www.clips.ua.ac.be/*

[20] AlchemyAPI, a product of Orchestr8, semantic tagging and text mining solutions, from: *http://www.alchemyapi.com*

20